STAT 506: Final Exam
Name:

Please turn in the exam to GitHub and include the R Markdown code and a PDF file with output. Please verify that all of the code has compiled and the graphics look like you think they should on your files, you are welcome to upload image files directly if they look distorted in the output file.

While the exam is open book and you can use any resources from class or freely available on the internet, this is strictly an individual endeavor and **you should not discuss the problems with anyone outside the course instructor including group mates or class members.** All resources, including websites, should be acknowledged.

For full credit, include your code and graphics for each question and create neat output by using options like `kable()` for tables and writing results in line with R commands. R output alone is not a sufficient answer and should be accompanied with a text narrative.

Question 1. (34 points)

A common approach to analyzing count data is to use the log counts in a regression model as opposed to using a probability distribution, such as Poisson or negative binomial, that is appropriate for the count response. The question will compare these two approaches.

A. (14 points)

i (4 points) Simulate count data from a negative binomial model regression model with a single continuous predictor variable.

ii (4 points) Create a plot to show the relationship between your predictor variable and the response.

iii (2 points) Write out the density function you selected for simulating from the negative binomial distribution.

iv (4 points) Now write out the GLM structure along with values for the parameters that you are using to simulate the data in the figure in A(ii).

B. (8 points)

You are interested in a mean-only model. Assume that for this process you believe the mean is $exp(5)$ and the size parameter is $\frac{1}{2}$ (which is parameterized below).

```
y <- rnbinom(n, mu = exp(5), size = .5)
```

i (6 points) Use simulation to estimate (approximately) how many samples are required to have a .5 probability that the standard error of the mean is less than 0.3.

ii. (2 points) Justify your answer and explain your approach.

C. (12 points)

This question will compare a negative binomial regression model with the log counts approach using the simulated data below.

i. (2 points) Fit a negative binomial regression model and compare the parameter estimates with the known values.

ii. (2 points) Fit a regression model on the log counts and compare the parameter estimates with the known values.

iii. (2 points) Compare the results from parts i and ii.

iv. (2 points) Using your model from i and a posterior predictive distribution compare the results with the observed data.

v. (2 points) Using your model from ii and a posterior predictive distribution compare the results with the observed data.

vi. (2 points) Compare the results from parts iv and v.

Question 2 (36 points)

Recall the lunge dataset used for project 3.

A. (12 points)

For this question use the lunge dataset and the model specified in the code below

```
lunge_final <- read_csv('lunge_final.csv')
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   subject = col_double(),
##   replicate = col_double(),
##   lunge_type = col_character(),
##   acceleration = col_double(),
##   muscle = col_double()
## )
```

```
lm(acceleration ~ muscle * lunge_type, data = lunge_final)
```

i. (2 points) Define the design matrix (X). Note there should be more than one column for categorical variables.

ii. (2 points) Write out the statistical model implied by this code.

iii. (4 points) The `lm()` function uses least squares to estimate the regression coefficients. Derive the OLS estimator of $\underline{\beta}$ and calculate this directly in R (using matrix algebra, not `lm` directly).

iv. (4 points) Summarize the output and describe what each coefficient means in the context of lunges and acceleration.

```
display(lm(acceleration ~ muscle * lunge_type, data = lunge_final))
```

```
## lm(formula = acceleration ~ muscle * lunge_type, data = lunge_final)
##                                 coef.est coef.se
## (Intercept)                      0.20     0.07
## muscle                           0.54     0.14
## lunge_typeLunge_Bosu_down        0.03     0.09
## lunge_typeLunge_Bosu_up          0.02     0.08
## lunge_typeLunge_Foam             0.01     0.09
## muscle:lunge_typeLunge_Bosu_down -0.05     0.18
## muscle:lunge_typeLunge_Bosu_up   -0.01     0.17
## muscle:lunge_typeLunge_Foam      -0.03     0.18
## ---
## n = 120, k = 8
## residual sd = 0.10, R-Squared = 0.41
```

B. (14 points)

i. (2 points) Question 2A ignored the repeated measurement on the athletes, why is this problematic?

ii. (2 points) Write out the statistical model implied by this code.

```
stan_glmer(acceleration ~ muscle * lunge_type + (1 | subject), data = lunge_final)
```

iii. (4 points) Does the statistical model in Question 2B(ii) address the issue of repeated measurements on athletes? If so, how?

iv. (2 points) Interpret the output of this code in the context of the model in Question 2B(ii)

```
print(stan_glmer(acceleration ~ muscle * lunge_type + (1 | subject), data = lunge_final))
```

```
## stan_glmer
##  family:       gaussian [identity]
##  formula:      acceleration ~ muscle * lunge_type + (1 | subject)
##  observations: 120
## ------
##                                 Median MAD_SD
## (Intercept)                      0.5    0.0
## muscle                          -0.1    0.1
## lunge_typeLunge_Bosu_down       -0.1    0.0
## lunge_typeLunge_Bosu_up         -0.1    0.0
## lunge_typeLunge_Foam             0.0    0.0
## muscle:lunge_typeLunge_Bosu_down 0.2    0.1
## muscle:lunge_typeLunge_Bosu_up   0.2    0.1
```

```
## muscle:lunge_typeLunge_Foam       0.0    0.1
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 0.0    0.0
##
## Error terms:
##  Groups   Name        Std.Dev.
##  subject  (Intercept) 0.114
##  Residual             0.031
## Num. levels: subject 10
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

v. (2 points) How does the output in Question 2B(iv) differ from that in Question 2A(iv)?

vi. (2 points) Suppose that age is a meaningful factor in an subject's ability to perform lunges. Modify the statistical model in Question 2B(ii) to allow for a covariate on the subject and write this out. Note you do not need to fit this model.

C. (10 points)

Perform causal inference using the lunge data with the goal of comparing muscle contraction across different lunge types (treatments). For this analysis the replicates are averaged for each lunge type so there is a single measurement for each subject and lunge type. We can treat the subjects as blocking variables, where each experimental unit is a time slot for an individual that has a lunge type applied to it.

```
lunge_causal <- lunge_final %>% group_by(subject, lunge_type) %>%
  summarize(mean_muscle = mean(muscle), .groups = 'drop')
```

- Write out and justify a model that you will use to investigate the differences in the treatments
- Fit model and summarize results. Focus on estimated causal effects on the four different treatments.